Pontifícia Universidade Católica
do Rio de Janeiro

Eraldo Luís Rezende Fernandes

Entropy Guided Feature Generation for Structure Learning

Tese de Doutorado

Thesis presented to the Programa de Pós-Graduação em Informática, of the Departamento de Informática do Centro Técnico Científico da PUC–Rio, as partial fulfillment of the requirements for the degree of Doutor.

Advisor: Prof. Ruy Luiz Milidiú

Rio de Janeiro
September 2012

Pontifícia Universidade Católica do Rio de Janeiro

Eraldo Luís Rezende Fernandes

Entropy Guided Feature Generation for Structure Learning

Thesis presented to the Programa de Pós-Graduação em Informática, of the Departamento de Informática do Centro Técnico Científico da PUC–Rio, as partial fulfillment of the requirements for the degree of Doutor.

Prof. Ruy Luiz Milidiú
Advisor
Departamento de Informática — PUC–Rio

Prof. Marcus Vinicius Soledade Poggi de Aragão
Departamento de Informática — PUC–Rio

Prof. Valmir Carneiro Barbosa
UFRJ

Prof. Cícero Nogueira dos Santos
IBM Research

Prof. Daniel Schwabe
Departamento de Informática — PUC–Rio

Prof. José Eugenio Leal
Coordinator of the Centro Técnico Científico — PUC–Rio

Rio de Janeiro — September 6, 2012

Eraldo Luís Rezende Fernandes

Graduated from the Universidade Federal de Mato Grosso do Sul in Ciência da Computação and obtained a degree of Mestre em Informática at PUC–Rio. He is a lecturer and a researcher at the Instituto Federal de Educação, Ciência e Tecnologia de Goiás.

Bibliographic data

Fernandes, Eraldo Luís Rezende

Entropy Guided Feature Generation for Structure Learning / Eraldo Luís Rezende Fernandes ; advisor: Ruy Luiz Milidiú. — 2012.
93 f. : il. ; 30 cm

Tese (Doutorado em Informática)-Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2012.
Inclui bibliografia

1. Informática – Teses. 2. Aprendizado de Estruturas. 3. Geração de Atributos. 4. Entropia. 5. Processamento de Linguagem Natural. I. Milidiú, Ruy Luiz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

First of all, I am thankful for all my family which has always supported me. We have had amazing moments during so many meetings, and it is always my pleasure to be with them. Specially, I would like to thank my parents, Eraldo and Sônia, my sister, Luciana, and my little brother, Lúciu. They constitute my ultimate foundation for all aspects of my life.

I am unable to express in words all my love and gratitude for my dear wife, Valéria. She standed by me during the most difficult and sad moments of my PhD course. And, more importantly, we are always enjoying each other and having fun with the simplest things.

I also want to thank my sisters-in-law, Vanessa and Patrícia, and my brother-in-law, Márcio.

I will never forget the great moments I have spent in the lab. Thank you very much, all *Learnlings*. Specially, I am glad for having closely worked with Carlos, Cícero, Eduardo, Leandro and William.

I am thankful to my advisor, Ruy, who have taught me so many things regarding different aspects of my life.

I thank all my friends from Barcelona, where I spent one great year at Yahoo! Research Lab. Specially, I need to thank my supervisor and friend, Ulf, and all the *Berlineros*.

Finally, I need to thank my colleagues from IFG for all the support. Specially, José Antônio, his family and Sérgio Henrique.

Abstract

Fernandes, Eraldo Luís Rezende; Milidiú, Ruy Luiz. **Entropy Guided Feature Generation for Structure Learning**. Rio de Janeiro, 2012. 93p. DSc Thesis — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Structure learning consists in learning a mapping from inputs to structured outputs by means of a sample of correct input-output pairs. Many important problems fit into this setting. Natural language processing provides several tasks that can be formulated and solved as structure learning problems. Dependency parsing, for instance, involves the prediction of a tree underlying a sentence. Feature generation is an important subtask of structure learning which, usually, is partially solved by a domain expert that builds complex discriminative feature templates by conjoining the available basic features. This is a limited and expensive way to generate features and is recognized as a modeling bottleneck.

In this work, we propose an automatic feature generation method for structure learning problems. This method is entropy guided since it generates complex features based on the conditional entropy of local output variables given the available input features. We experimentally compare the proposed method with two important alternative feature generation methods, namely manual template generation and polynomial kernel methods. Our experimental findings indicate that the proposed method is more attractive than both alternatives. It is much cheaper than manual templates and computationally faster than kernel methods. Additionally, it is simpler to control its generalization performance than with kernel methods.

We evaluate our method on nine datasets involving five natural language processing tasks and four languages. The resulting systems present state-of-the-art comparable performances and, particularly on part-of-speech tagging, text chunking, quotation extraction and coreference resolution, remarkably achieve the best known performances on different languages like Arabic, Chinese, English, and Portuguese. Furthermore, our coreference resolution systems achieve the very first place on the Conference on Computational Natural Language Learning 2012 Shared Task. The competing systems were ranked by the mean score over three languages: Arabic, Chinese and English. Our approach obtained the best performances among all competitors for all the three languages.

Our feature generation method naturally extends the general structure learning framework and is not restricted to natural language processing tasks.

Keywords

Structure Learning. Feature Generation. Entropy. Natural Language Processing.

Resumo

Fernandes, Eraldo Luís Rezende; Milidiú, Ruy Luiz. **Geração de Atributos Guiada por Entropia para Aprendizado de Estruturas**. Rio de Janeiro, 2012. 93p. Tese de Doutorado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Aprendizado de estruturas consiste em aprender um mapeamento de variáveis de entrada para saídas estruturadas a partir de exemplos de pares entrada-saída. Vários problemas importantes podem ser modelados desta maneira. O processamento de linguagem natural provê diversas tarefas que podem ser formuladas e solucionadas através do aprendizado de estruturas. Por exemplo, parsing de dependência envolve o reconhecimento de uma *árvore* implícita em uma frase. Geração de atributos é uma sub-tarefa importante do aprendizado de estruturas. Geralmente, esta sub-tarefa é realizada por um especialista que constrói gabaritos de atributos complexos e discriminativos através da combinação dos atributos básicos disponíveis na entrada. Esta é uma forma limitada e cara para geração de atributos e é reconhecida como um gargalo de modelagem.

Neste trabalho, propomos um método automático para geração de atributos para problemas de aprendizado de estruturas. Este método é *guiado por entropia* já que é baseado na entropia condicional de variáveis locais de saída dados os atributos básicos. Comparamos experimentalmente o método proposto com dois métodos alternativos para geração de atributos: geração manual e métodos de kernel polinomial. Nossos resultados mostram que o método de geração de atributos guiado por entropia é superior aos dois métodos alternativos em diferentes aspectos. Nosso método é muito mais barato do que o método manual e computacionalmente mais rápido que o método baseado em kernel. Adicionalmente, ele permite o controle do seu poder de generalização mais facilmente do que métodos de kernel.

Nós avaliamos nosso método em nove datasets envolvendo cinco tarefas de linguística computacional e quatro idiomas. Os sistemas desenvolvidos apresentam resultados comparáveis aos melhores sistemas atualmente e, particularmente para etiquetagem morfossintática, identificação de sintagmas, extração de citações e resolução de coreferência, obtêm os melhores resultados conhecidos para diferentes idiomas como Árabe, Chinês, Inglês e Português. Adicionalmente, nosso sistema de resolução de coreferência obteve o primeiro lugar na competição *Conference on Computational Natural Language Learning 2012 Shared Task*. O sistema vencedor foi determinado pela média de desempenho em três idiomas: Árabe, Chinês e Inglês. Nosso sistema obteve o melhor desempenho nos três idiomas avaliados.

Nosso método de geração de atributos estende naturalmente o framework de aprendizado de estruturas e não está restrito a tarefas de processamento de linguagem natural.

Palavras–chave

Aprendizado de Estruturas. Geração de Atributos. Entropia. Processamento de Linguagem Natural.

Contents

List of Figures

List of Tables